

Exact ABC using Importance Sampling

Minh-Ngoc Tran and Robert Kohn *

September 29, 2015

Abstract

Approximate Bayesian Computation (ABC) is a powerful method for carrying out Bayesian inference when the likelihood is computationally intractable. However, a drawback of ABC is that it is an approximate method that induces a systematic error because it is necessary to set a tolerance level to make the computation tractable. The issue of how to optimally set this tolerance level has been the subject of extensive research. This paper proposes an ABC algorithm based on importance sampling that estimates expectations with respect to the *exact* posterior distribution given the observed summary statistics. This overcomes the need to select the tolerance level. By *exact* we mean that there is no systematic error and the Monte Carlo error can be made arbitrarily small by increasing the number of importance samples. We provide a formal justification for the method and study its convergence properties. The method is illustrated in two applications and the empirical results suggest that the proposed ABC based estimators consistently converge to the true values as the number of importance samples increases. Our proposed approach can be applied more generally to any importance sampling problem where an unbiased estimate of the likelihood is required.

Keywords. Approximate Bayesian Computation, Debiasing, Ising model, Marginal likelihood Estimate , Unbiased likelihood Estimate

*Minh-Ngoc Tran is with the Business Analytics discipline, University of Sydney Business School, Sydney 2006 Australia (minh-ngoc.tran@sydney.edu.au). Robert Kohn is with the UNSW Business School, University of New South Wales, Sydney 2052 Australia (r.kohn@unsw.edu.au).

1 Introduction

Many Bayesian inference problems, including the calculation of posterior moments and probabilities, require evaluating an integral of the form

$$\mathbb{E}(\varphi|y_{\text{obs}}) = \int_{\Theta} \varphi(\theta)p(\theta|y_{\text{obs}})d\theta. \quad (1)$$

In (1), $p(\theta|y_{\text{obs}}) \propto p(\theta)p(y_{\text{obs}}|\theta)$ is the posterior distribution of θ , y_{obs} is the observed data, $\theta \in \Theta$ is the vector of model parameters, and $\varphi(\theta)$ is a function mapping Θ to the real line. In many problems the likelihood $p(y|\theta)$ is intractable, either because it cannot be computed or because it is too expensive to compute. The Approximate Bayesian Computation (ABC) approach was proposed to overcome this problem as it only requires that we are able to sample from the model density $y \sim p(\cdot|\theta)$ without being able to evaluate it (see Tavaré et al., 1997; Beaumont et al., 2002; Marjoram et al., 2003; Sisson and Fan, 2011).

ABC approximates the intractable likelihood $p(y_{\text{obs}}|\theta)$ by

$$p_{\text{ABC},\epsilon}(y_{\text{obs}}|\theta) := \int K_{\epsilon}(y - y_{\text{obs}})p(y|\theta)dy \quad (2)$$

with $K_{\epsilon}(u)$ a scaled kernel density with bandwidth $\epsilon > 0$. If the original dataset y has a complex structure and is high dimensional, it is computationally more efficient and convenient to work with a lower-dimensional summary statistic $s = S(y) \in \mathbb{R}^d$. That is, instead of (2), we work with

$$p_{\text{ABC},\epsilon}(s_{\text{obs}}|\theta) := \int K_{\epsilon}(s - s_{\text{obs}})p(s|\theta)ds, \quad (3)$$

where $p(s|\theta)$ denotes the density of summary statistic s and $s_{\text{obs}} = S(y_{\text{obs}})$. Here, $K_{\epsilon}(u) = K(u/\epsilon)/\epsilon^d$ with $K(\cdot)$ a d -variate kernel density such as a Gaussian density.

There are two major approximations used in ABC. The first is using a summary statistic instead of the original data

$$p(\theta|y_{\text{obs}}) \approx p(\theta|s_{\text{obs}}) \propto p(\theta)p(s_{\text{obs}}|\theta), \quad (4)$$

which is exact if the summary statistic $S(\cdot)$ is sufficient. The second results from approximating the intractable likelihood $p(s_{\text{obs}}|\theta)$ by $p_{\text{ABC},\epsilon}(s_{\text{obs}}|\theta)$,

$$p(s_{\text{obs}}|\theta) \approx p_{\text{ABC},\epsilon}(s_{\text{obs}}|\theta) = \int K_{\epsilon}(s - s_{\text{obs}})p(s|\theta)ds. \quad (5)$$

This approximation is exact if $\epsilon = 0$, but setting $\epsilon = 0$ is impractical as the event that $s = s_{\text{obs}}$ occurs with probability zero for all but the simplest applications. Selecting ϵ is still an open question because it is usually necessary to trade off between computational load and accuracy when selecting ϵ .

All current ABC algorithms suffer from approximation errors due to approximation (4), if $S(\cdot)$ is not sufficient, and approximation (5) if $\epsilon > 0$. Our article proposes an ABC algorithm to estimate (1) that completely removes the error due to approximation (5), i.e. we are able to estimate expectations with respect to the *exact* posterior $p(\theta|s_{\text{obs}})$ based on the summary statistic. In addition, if $S(\cdot)$ is sufficient, then the estimated expectations are with respect to the exact posterior $p(\theta|y_{\text{obs}})$.

The basic idea is to obtain an unbiased estimator of the likelihood, based on the debiasing approach of McLeish (2012) and Rhee and Glynn (2013). We then construct an importance sampling estimator of the integral (1) and establish its convergence properties. The unbiasedness allows the importance sampling estimator to converge almost surely to the true value (1) without suffering from the systematic error associated with the use of $\epsilon > 0$. We illustrate the proposed method by a Gaussian example and an Ising model example.

We note that our approach can be applied more generally to importance sampling problems where an unbiased estimate of the likelihood is required.

2 Constructing an unbiased estimator using a debiasing approach

Let λ be an unknown constant that we want to estimate and let ζ_k , $k=0,1,\dots$ be a sequence of biased estimators of λ , such that it is possible to generate ζ_k for each k . We are interested in constructing an unbiased estimator $\hat{\lambda}$ of λ , i.e. $\mathbb{E}(\hat{\lambda}) = \lambda$, based on the ζ_k 's, so that $\hat{\lambda}$ has a finite variance. We now present the debiasing approach, proposed independently by McLeish (2012) and Rhee and Glynn (2013), for constructing such a $\hat{\lambda}$. The basic idea is to introduce randomization into the sequence $\{\zeta_k, k=0,1,2,\dots\}$ to eliminate the bias.

Proposition 1 (Theorem 1 of Rhee and Glynn (2013)). *Suppose that T is a non-negative integer-valued random variable such that $P(T \geq k) > 0$ for any $k = 0,1,2,\dots$, and that T is independent of the ζ_k 's. Let $\varpi_k := 1/P(T \geq k)$. If*

$$\sum_{k=1}^{\infty} \varpi_k \mathbb{E}((\zeta_{k-1} - \lambda)^2) < \infty, \quad (6)$$

then

$$\hat{\lambda} := \zeta_0 + \sum_{k=1}^T \varpi_k (\zeta_k - \zeta_{k-1}),$$

is an unbiased estimator of λ and has the finite variance

$$\mathbb{V}(\hat{\lambda}) = \sum_{k=1}^{\infty} \varpi_k \left(\mathbb{E}((\zeta_{k-1} - \lambda)^2) - \mathbb{E}((\zeta_k - \lambda)^2) \right) - \mathbb{E}((\zeta_0 - \lambda)^2) < \infty. \quad (7)$$

3 Exact ABC

3.1 Constructing an unbiased estimator of the likelihood

Let $\epsilon_k, k=0,1,\dots$ be a sequence of monotonically decreasing positive numbers and n_k a sequence of monotonically increasing positive integers such that $\epsilon_k \rightarrow 0$ and $n_k \rightarrow \infty$ as $k \rightarrow \infty$. We estimate the ABC likelihood $p_{\text{ABC},\epsilon_k}(s_{\text{obs}}|\theta)$ based on the n_k pseudo-datasets $s_i^k \sim p(\cdot|\theta)$, $i=1,\dots,n_k$, as

$$\zeta_k := \widehat{p}_{\text{ABC},\epsilon_k}(s_{\text{obs}}|\theta) = \frac{1}{n_k} \sum_{i=1}^{n_k} K_{\epsilon_k}(s_i^k - s_{\text{obs}}). \quad (8)$$

It is important to note that the pseudo datasets s_i^k , $i=1,\dots,n_k$, can be re-used to compute ζ_j with $j > k$ as it is unnecessary that the ζ_k 's in Proposition 1 are independent. This significantly reduces the computational cost when it is expensive to generate these pseudo-datasets from $s \sim p(\cdot|\theta)$.

Theorem 1. *Let $K(\cdot)$ be a d -multivariate kernel density, i.e. $K(x) \geq 0$, $\int K(x)dx = 1$. We assume that*

$$\int xK(x)dx = 0, \sigma_K^2 := \int x'xK(x)dx < \infty, \sigma_R^2 := \int K^2(x)dx < \infty, \int x'xK^2(x)dx < \infty. \quad (9)$$

Let T be a non-negative integer-valued random variable that is independent of the $\zeta_k, k \geq 0$, and such that $P(T \geq k) > 0$ for any $k \geq 0$. Let $\varpi_k = 1/P(T \geq k)$. Suppose that $p(s|\theta)$ is twice differentiable in s for every θ , and that

$$\sum_{k=1}^{\infty} \varpi_k \left(\epsilon_{k-1}^4 + \frac{1}{n_{k-1} \epsilon_{k-1}^d} \right) < \infty. \quad (10)$$

Then,

$$\widehat{p}(s_{\text{obs}}|\theta) := \zeta_0 + \sum_{k=1}^T \varpi_k (\zeta_k - \zeta_{k-1}),$$

is an unbiased estimator of $p(s_{\text{obs}}|\theta)$ and has a finite variance.

We now use Theorem 1 to construct an unbiased estimator $\widehat{p}(s_{\text{obs}}|\theta)$ of the posterior $p(s_{\text{obs}}|\theta)$ by designing T , ϵ_k and n_k to satisfy the conditions of Theorem 1. Let T be a non-negative integer-valued random variable such that $P(T = k) := \rho(1-\rho)^k$, $k = 0,1,\dots$ for $0 < \rho < 1$. This choice means that the closer ρ is to 0, the bigger the values that T is likely to take. Then $\varpi_k = 1/P(T \geq k) = 1/(1-\rho)^k$. Let τ be a number such that $0 < \tau < 1$. If we select

$$\epsilon_k := [\tau(1-\rho)]^{\frac{k+1}{4}} \quad \text{and} \quad n_k := \left\lceil \frac{1}{[\tau(1-\rho)]^{(k+1)(1+d/4)}} \right\rceil,$$

then

$$\sum_{k=1}^{\infty} \varpi_k \left(\epsilon_{k-1}^4 + \frac{1}{n_{k-1} \epsilon_{k-1}^d} \right) < 2 \sum_{k=1}^{\infty} \tau^k < \infty.$$

That is, condition (10) is satisfied.

From (7) and (16), after some algebra, the variance $\mathbb{V}(\hat{p}(s_{\text{obs}}|\theta))$ is approximately

$$\mathbb{V}(\hat{p}(s_{\text{obs}}|\theta)) \approx (C_1 + C_2) \left(1 - \tau(1 - \rho) \right) \frac{\tau}{1 - \tau} - (C_1 + C_2) (\tau(1 - \rho))^{1/4},$$

with C_1 and C_2 positive constants in the proof of Theorem 1. The first term, which dominates the second term, is a monotonic increasing function of τ and ρ . So the variance $\mathbb{V}(\hat{p}(s_{\text{obs}}|\theta))$ will be small if ρ and τ are close to 0. However, small ρ and τ lead to a large k and hence a large n_k , especially if d is large. We can reduce the variance of the unbiased estimator $\hat{p}(s_{\text{obs}}|\theta)$ by using $\overline{\hat{p}(s_{\text{obs}}|\theta)} = (\hat{p}(s_{\text{obs}}|\theta)_1 + \dots + \hat{p}(s_{\text{obs}}|\theta)_{n_{\text{rep}}}) / n_{\text{rep}}$, with the $\hat{p}(s_{\text{obs}}|\theta)_i$ independent replications of $\hat{p}(s_{\text{obs}}|\theta)$. Then $\mathbb{E}(\overline{\hat{p}(s_{\text{obs}}|\theta)}) = p(s_{\text{obs}}|\theta)$. This approach to estimating $p(s_{\text{obs}}|\theta)$ has the important advantage that it automatically gives an estimate of $\mathbb{V}(\hat{p}(s_{\text{obs}}|\theta))$ and hence $\mathbb{V}(\overline{\hat{p}(s_{\text{obs}}|\theta)}) = \mathbb{V}(\hat{p}(s_{\text{obs}}|\theta)) / n_{\text{rep}}$, i.e.,

$$\hat{\mathbb{V}}(\hat{p}(s_{\text{obs}}|\theta)) = \frac{\sum_{i=1}^{n_{\text{rep}}} \left(\hat{p}(s_{\text{obs}}|\theta)_i - \overline{\hat{p}(s_{\text{obs}}|\theta)} \right)^2}{(n_{\text{rep}} - 1)} \quad \text{and} \quad \hat{\mathbb{V}}(\overline{\hat{p}(s_{\text{obs}}|\theta)}) = \frac{\hat{\mathbb{V}}(\hat{p}(s_{\text{obs}}|\theta))}{n_{\text{rep}}}.$$

3.2 Exact ABC with IS²

Define $\pi(\theta) := p(\theta|s_{\text{obs}})$ and let $\hat{p}(s_{\text{obs}}|\theta, u)$ be the unbiased estimator of $p(s_{\text{obs}}|\theta)$ obtained using the debiasing approach described in the previous section, and $u \in \mathcal{U}$ is the set of uniform random variables used to generate T and ζ_0, \dots, ζ_T . We denote by $p(u|\theta, s_{\text{obs}})$ the density of u and sometimes write $p(u|\theta, s_{\text{obs}})$ as $p(u|\theta)$ for notational simplicity. If the unbiased estimator $\hat{p}(s_{\text{obs}}|\theta, u)$ is non-negative almost surely for each θ , then we could use the pseudo-marginal Metropolis-Hastings (PMMH) algorithm (Andrieu and Roberts, 2009) to sample from the posterior $p(\theta|s_{\text{obs}})$. In general, however, the debiased estimator $\hat{p}(s_{\text{obs}}|\theta, u)$ can be negative, so it is mathematically invalid to use PMMH in our situation. See Jacob and Thiery (2015) for a detailed discussion.

Suppose that we wish to estimate the expectation of the function $\varphi(\theta)$ on Θ with respect to the posterior distribution, i.e.,

$$\mathbb{E}_{\pi}(\varphi) = \int_{\Theta} \varphi(\theta) \pi(\theta) d\theta = \frac{\int_{\Theta} \varphi(\theta) p(s_{\text{obs}}|\theta) p(\theta) d\theta}{\int_{\Theta} p(s_{\text{obs}}|\theta) p(\theta) d\theta}.$$

Then,

$$\mathbb{E}_{\pi}(\varphi) = \frac{\int_{\Theta} \int_{\mathcal{U}} \varphi(\theta) \hat{p}(s_{\text{obs}}|\theta, u) p(\theta) p(u|\theta, s_{\text{obs}}) d\theta du}{\int_{\Theta} \int_{\mathcal{U}} \hat{p}(s_{\text{obs}}|\theta, u) p(\theta) p(u|\theta, s_{\text{obs}}) d\theta du}.$$

Let $g_{\text{IS}}(\theta)$ be an importance density on Θ . For a function $h(\theta)$ of $\theta \in \Theta$, define

$$I(h) := \int_{\Theta} h(\theta) p(s_{\text{obs}}|\theta) p(\theta) d\theta = \int_{\Theta} \int_{\mathcal{U}} h(\theta) \frac{\widehat{p}(s_{\text{obs}}|\theta, u) p(\theta)}{g_{\text{IS}}(\theta)} g_{\text{IS}}(\theta) p(u|\theta, s_{\text{obs}}) d\theta du$$

which is unbiasedly estimated by

$$\widehat{I}(h) := \frac{1}{M} \sum_{i=1}^M h(\theta_i) \widehat{w}(\theta_i, u_i),$$

where

$$\theta_i \sim g_{\text{IS}}(\cdot), u_i \sim p(\cdot|\theta_i, s_{\text{obs}}) \quad \text{and} \quad \widehat{w}(\theta_i, u_i) := \frac{\widehat{p}(s_{\text{obs}}|\theta_i, u_i) p(\theta_i)}{g_{\text{IS}}(\theta_i)}. \quad (11)$$

We now define the estimate of $\mathbb{E}_{\pi}(\varphi)$ as

$$\widehat{\mathbb{E}_{\pi}(\varphi)} := \frac{\widehat{I}(\varphi)}{\widehat{I}(1)}. \quad (12)$$

In this form, the estimator $\widehat{\mathbb{E}_{\pi}(\varphi)}$ is similar to the IS^2 estimator introduced in Tran et al. (2013), who propose an importance sampling procedure when the likelihood is intractable but a non-negative unbiased estimator of the likelihood is available.

We now summarize the algorithm for estimating $\mathbb{E}_{\pi}(\varphi)$, and refer to it as the Exact ABC algorithm based on an IS^2 approach, or EABC- IS^2 for short.

Algorithm 1 (EABC- IS^2 algorithm). *For $i=1, \dots, M$*

- *Generate $\theta_i \sim g_{\text{IS}}(\cdot)$, $u_i \sim p(\cdot|\theta_i, s_{\text{obs}})$ and compute $\widehat{p}(s_{\text{obs}}|\theta_i, u_i)$.*
- *Compute the weights $\widehat{w}(\theta_i, u_i)$ as in (11).*

Compute the EABC- IS^2 estimator $\widehat{\mathbb{E}_{\pi}(\varphi)}$ of $\mathbb{E}_{\pi}(\varphi)$ as in (12).

Remark 1. *As with all importance sampling, it is straightforward to estimate several expectations simultaneously at almost the same cost as one expectation, because the weights $\widehat{w}(\theta_i, u_i)$ are the same.*

To obtain a strong law of large numbers and a central limit theorem for $\widehat{\mathbb{E}_{\pi}(\varphi)}$ we define $\xi(\theta, u) := \widehat{p}(s_{\text{obs}}|\theta, u)/p(s_{\text{obs}}|\theta)$, so that $\mathbb{E}_{u \sim p(\cdot|\theta)}(\xi(\theta, u)) = 1$.

Theorem 2. *Suppose that $\text{Sup}(\pi) \subseteq \text{Sup}(g_{\text{IS}})$, where Sup means support.*

- (i) *If $\mathbb{E}_{\pi}(|\varphi(\theta)|) < \infty$, then $\widehat{\mathbb{E}_{\pi}(\varphi)} \rightarrow \mathbb{E}_{\pi}(\varphi)$ almost surely as $M \rightarrow \infty$.*

(ii) If $\mathbb{E}_{g_{\text{IS}}} \left(\frac{\mathbb{E}_{u \sim p(\cdot|\theta)} (\xi^2(\theta, u)) \varphi(\theta)^2 \pi(\theta)^2}{g_{\text{IS}}^2(\theta)} \right) < \infty$ then $\sqrt{M} \left(\widehat{\mathbb{E}_\pi(\varphi)} - \mathbb{E}_\pi(\varphi) \right) \rightarrow \mathcal{N}(0, \sigma_\varphi^2)$ as $M \rightarrow \infty$, where

$$\sigma_\varphi^2 := \mathbb{E}_{g_{\text{IS}}} \left(\frac{\pi^2(\theta)}{g_{\text{IS}}^2(\theta)} (\varphi(\theta) - \mathbb{E}_\pi(\varphi))^2 \mathbb{E}_{u \sim p(\cdot|\theta)} (\xi^2(\theta, u)) \right). \quad (13)$$

If we can evaluate $p(s_{\text{obs}}|\theta)$ so that $\xi = 1$, then $\sigma_\varphi^2 = \mathbb{E}_{g_{\text{IS}}} \left(\frac{\pi^2(\theta)}{g_{\text{IS}}^2(\theta)} (\varphi(\theta) - \mathbb{E}_\pi(\varphi))^2 \right)$ is the variance of the noiseless importance sampler.

(iii) $\widehat{\sigma_\varphi^2}$ is a consistent estimator of σ_φ^2 , where

$$\widehat{\sigma_\varphi^2} := \frac{1}{M \widehat{p}(s_{\text{obs}})^2} \sum_{i=1}^M (\varphi(\theta_i) - \widehat{\mathbb{E}_\pi(\varphi)})^2 \widehat{w}^2(\theta_i, u_i),$$

and

$$\widehat{p}(s_{\text{obs}}) := \frac{1}{M} \sum_{i=1}^M \widehat{w}(\theta_i, u_i). \quad (14)$$

Remark 2. We note that $\widehat{p}(s_{\text{obs}})$ in (14) is an estimate of the marginal likelihood $p(s_{\text{obs}})$, which can be used for model comparison. It is straightforward to obtain this marginal likelihood estimate and an estimate of its standard error and we can readily show that $\widehat{p}(s_{\text{obs}})$ converges to $p(s_{\text{obs}})$ as $M \rightarrow \infty$. It is usually difficult to accurately estimate the marginal likelihood and its standard error using competing ABC approaches.

4 Examples

4.1 A Gaussian example

This example is discussed by Sisson and Fan (2011) who consider a univariate Gaussian model $y \sim \mathcal{N}(\theta, 1)$, with $y_{\text{obs}} = 0$ and a diffuse prior $p(\theta) \propto 1$. Here, the posterior is $\pi(\theta) = p(\theta|y_{\text{obs}}) = \mathcal{N}(0, 1)$ and the summary statistics $s = S(y) = y$ is sufficient. We are interested in estimating the posterior noncentral second moment of θ ,

$$\mathbb{E}(\theta^2|y_{\text{obs}}) = \int \theta^2 p(\theta|y_{\text{obs}}) d\theta = 1.$$

We take the kernel $K(\cdot)$ as the standard normal density, so the ABC likelihood $p_{\text{ABC}, \epsilon}(y_{\text{obs}}|\theta)$ in (3) can be computed analytically, and the ABC posterior is $p_{\text{ABC}, \epsilon}(\theta|y_{\text{obs}}) \propto p(\theta) p_{\text{ABC}, \epsilon}(y_{\text{obs}}|\theta) = \mathcal{N}(0, 1 + \epsilon^2)$. So setting aside the Monte Carlo error, standard ABC procedures estimate $\mathbb{E}(\theta^2|y_{\text{obs}})$ by $1 + \epsilon^2$, which always suffers from a systematic error whenever $\epsilon > 0$.

M	1000	10,000	100,000	1,000,000
EABC-IS ² estimate	1.0065 (0.0733)	1.0044 (.0245)	1.0008 (0.0111)	1.0000 (0.0002)

Table 1: EABC-IS² estimates of $\mathbb{E}(\theta^2|y_{\text{obs}})=1$ for various numbers of samples M . The numbers in brackets are standard errors

To run the EABC-IS² algorithm, we estimate the likelihood $p(y_{\text{obs}}|\theta)$ unbiasedly using the debiasing approach in Section 3.1 with $\rho=0.4$, $\tau=0.2$ and the importance density $g_{\text{IS}}(\theta) = \mathcal{N}(0,2)$. The number of replications n_{rep} is selected such that the variance $\mathbb{V}(\log|\hat{p}(y_{\text{obs}}|\bar{\theta})|) \approx 1$ with $\bar{\theta} = 0.5$. This is motivated by the IS² theory in Tran et al. (2013) who show that the optimal variance of the log-likelihood estimators is 1 in order to minimize the overall computational cost.

Table 1 shows the EABC-IS² estimates of $\mathbb{E}(\theta^2|y_{\text{obs}})$ for various numbers of samples M . The results suggest empirically that the estimates consistently get closer to the true value as M increases. This attractive property of the EABC-IS² is contrasted with other ABC algorithms where a systematic error always exists no matter how large M is.

4.2 Ising model

Our second example is the Ising model on a rectangular lattice of size $L \times W$ with data $y_{i,j} \in \{-1,1\}$ and likelihood

$$p(y|\theta) = \frac{\exp(\theta S(y))}{C(\theta)},$$

where $S(y) = \sum_{i=1}^{L-1} \sum_{j=1}^W y_{i,j} y_{i+1,j} + \sum_{i=1}^L \sum_{j=1}^{W-1} y_{i,j} y_{i,j+1}$; see Moller et al. (2006). The likelihood $p(y|\theta)$ has $S(y)$ as sufficient statistic and is considered intractable as computing the normalising constant $C(\theta)$ for each θ is infeasible for large lattices. However, one can generate data y from the Ising model $y \sim p(\cdot|\theta)$ using, for example, perfect simulation or Monte Carlo simulation. We note that $S(y)$ is a sufficient statistic for θ .

In this example, we set $L=W=50$ and generate a data set y_{obs} using $\theta=0.5$. Our task is to estimate the posterior mean of θ , given y_{obs} . As in Moller et al. (2006), we use a uniform prior $U(0,1)$ for θ . For this Ising model, an exact MCMC is available for sampling from the posterior $p(\theta|y_{\text{obs}})$ (Moller et al., 2006), which we use as the “gold standard” for comparison. We run this exact MCMC algorithm for 1,000,000 iterations and obtain an estimate of 0.5099. for the posterior mean $\mathbb{E}(\theta|y_{\text{obs}}) = \int \theta p(\theta|y_{\text{obs}}) d\theta$. The number in brackets is the standard deviation.

The EABC-IS² estimate of $\mathbb{E}(\theta|y_{\text{obs}})$, based on $M=200,000$ samples of θ , is 0.5099 (0.0001) which is equal to (up to 4 decimal places) the estimate given by the exact MCMC algorithm.

We now use PMMH to sample from the ABC posterior $p_{\text{ABC},\epsilon}(\theta|y_{\text{obs}})$, for various $\epsilon=10, 1$ and 0.1 , with the ABC likelihood $p_{\text{ABC},\epsilon}(y_{\text{obs}}|\theta)$ in (2) estimated unbiasedly by

$$\hat{p}_{\text{ABC},\epsilon}(s_{\text{obs}}|\theta) = \frac{1}{n} \sum_{i=1}^n K_{\epsilon}(s_i - s_{\text{obs}}), \quad s_i \sim p(\cdot|\theta).$$

For each ϵ , the number of pseudo datasets n is tailored such that the acceptance rate is about 0.23. The ABC-PMMH estimates of the posterior mean $\mathbb{E}_{\theta \sim p_{\text{ABC},\epsilon}(\theta|y_{\text{obs}})}(\theta|y_{\text{obs}})$, based on 100,000 iterations, are 0.5094 (0.0002), 0.5108 (0.0003) and 0.5100 (0.0003) respectively. These estimates get closer to the “gold standar” estimate 0.5099 when ϵ decreases. Note that the smaller the value of ϵ , the greater the computational cost as we need a bigger n in order for the Markov chain to mix well.

5 Discussion

Our article presents the EABC-IS² approach for estimating expectations with respect to the exact posterior distribution conditional on the observed summary statistic. The EABC-IS² estimators do not suffer from a systematic error inherent in standard ABC algorithms due to the use of tolerance $\epsilon > 0$. Our approach generalises directly to other applications of importance sampling where the likelihood is intractable but an unbiased estimator of the likelihood can be used.

Appendix: Proofs

Proof of Theorem 1. For a fixed θ , let $\lambda = p(s_{\text{obs}}|\theta)$. We first show that

$$\left(p_{\text{ABC},\epsilon_k}(y|\theta) - \lambda\right)^2 = \frac{1}{4}\epsilon_k^4 \sigma_K^4 \left(\text{tr}(\nabla^2 p(s_{\text{obs}}|\theta))\right)^2 + o(\epsilon_k^4). \quad (15)$$

$$\begin{aligned} p_{\text{ABC},\epsilon_k}(s_{\text{obs}}|\theta) &= \frac{1}{\epsilon_k^d} \int K\left(\frac{s - s_{\text{obs}}}{\epsilon_k}\right) p(s|\theta) ds \\ &= \int K(w) p(s_{\text{obs}} + \epsilon_k w|\theta) dw, \quad \text{where } w := \frac{s - s_{\text{obs}}}{\epsilon_k} \\ &= \int K(w) \left(p(s_{\text{obs}}|\theta) + \epsilon_k w' \nabla p(s_{\text{obs}}|\theta) + \frac{1}{2} \epsilon_k^2 w' \nabla^2 p(s_{\text{obs}}|\theta) w + o(\epsilon_k^2)\right) dw \\ &= p(s_{\text{obs}}|\theta) + \frac{1}{2} \epsilon_k^2 \sigma_K^2 \text{tr}(\nabla^2 p(s_{\text{obs}}|\theta)) + o(\epsilon_k^2), \end{aligned}$$

which gives (15). Similarly,

$$\mathbb{V}(\zeta_k) = n_k^{-1} \epsilon_k^{-d} R_K p(s_{\text{obs}}|\theta) + o(n_k^{-1} \epsilon_k^{-d}),$$

where $R_K = \int K(x)^2 dx$.

Then,

$$\begin{aligned}\mathbb{E}((\zeta_k - \lambda)^2) &= \mathbb{V}(\zeta_k) + (p_{\text{ABC}, \epsilon_k}(y|\theta) - \lambda)^2 \\ &= C_1 \epsilon_k^4 + C_2 n_k^{-1} \epsilon_k^{-d} + o(\epsilon_k^4 + n_k^{-1} \epsilon_k^{-d}),\end{aligned}\tag{16}$$

and (10) implies (6). The proof then follows from Proposition 1 \square

Proof of Theorem 2 . The proof is similar to that of Theorem 1 in Tran et al. (2013). Let $\tilde{g}_{\text{IS}}(\theta, u) := g_{\text{IS}}(\theta)p(u|\theta, s_{\text{obs}})$ and $\tilde{\pi}(\theta, u) := \pi(\theta)p(u|\theta, s_{\text{obs}})$. The condition $\text{Sup}(\pi) \subseteq \text{Sup}(g_{\text{IS}})$ implies that $\text{Sup}(\tilde{\pi}) \subseteq \text{Sup}(\tilde{g}_{\text{IS}})$. This, together with the existence and finiteness of $\mathbb{E}_\pi(\varphi)$ ensure that

$$\mathbb{E}_{\tilde{g}_{\text{IS}}}[\varphi(\theta_i)\widehat{w}(\theta_i, u_i)] = p(s_{\text{obs}})\mathbb{E}_\pi(\varphi) \quad \text{and} \quad \mathbb{E}_{\tilde{g}_{\text{IS}}}[\widehat{w}(\theta_i, u_i)] = p(s_{\text{obs}})$$

exist and are finite. Result (i) then follows immediately from (12) and the strong law of large numbers.

To prove (ii), write

$$\begin{aligned}\widehat{\mathbb{E}_\pi(\varphi)} - \mathbb{E}_\pi(\varphi) &= \frac{\frac{1}{M} \sum_{i=1}^M (\varphi(\theta_i) - \mathbb{E}_\pi(\varphi)) \widehat{w}(\theta_i, u_i)}{\frac{1}{M} \sum_{i=1}^M \widehat{w}(\theta_i, u_i)} = S_M / \widehat{p}(s_{\text{obs}}), \\ \text{where } S_M &= M^{-1} \sum_{i=1}^M X(\theta_i, u_i), \quad \text{with} \quad X(\theta, u) = (\varphi(\theta) - \mathbb{E}_\pi(\varphi)) \widehat{w}(\theta, u)\end{aligned}$$

The $X_i := X(\theta_i, u_i)$ are independently and identically distributed and it is straightforward to check that $\mathbb{E}_{\tilde{g}_{\text{IS}}}(X) = 0$.

$$\begin{aligned}\mathbb{V}_{\tilde{g}_{\text{IS}}}(X) &= \mathbb{E}_{\tilde{g}_{\text{IS}}}(X^2) \\ &= \mathbb{E}_{g_{\text{IS}}}(\mathbb{E}_{u \sim p(\cdot|\theta, s_{\text{obs}})}(X^2)) \\ &= \mathbb{E}_{g_{\text{IS}}}\left(\left((\varphi(\theta) - \mathbb{E}_\pi(\varphi)) \frac{p(\theta)p(s_{\text{obs}}|\theta)}{g_{\text{IS}}(\theta)}\right)^2 \mathbb{E}_{u \sim p(\cdot|\theta, s_{\text{obs}})}(\xi^2)\right) \\ &= p(s_{\text{obs}})^2 \mathbb{E}_{g_{\text{IS}}}\left(\left((\varphi(\theta) - \mathbb{E}_\pi(\varphi)) \frac{\pi(\theta)}{g_{\text{IS}}(\theta)}\right)^2 \mathbb{E}_{u \sim p(\cdot|\theta, s_{\text{obs}})}(\xi^2)\right) \\ &= p(s_{\text{obs}})^2 \sigma_\varphi^2\end{aligned}$$

By the central limit theorem for a sum of independently and identically distributed random variables with a finite second moment, $\sqrt{M}S_M \xrightarrow{d} \mathcal{N}(0, p(s_{\text{obs}})^2 \sigma_\varphi^2)$. By (i) and Slutsky's theorem,

$$\sqrt{M}(\widehat{\mathbb{E}_\pi(\varphi)} - \mathbb{E}_\pi(\varphi)) = \frac{\sqrt{M}S_M}{\widehat{p}(s_{\text{obs}})} \xrightarrow{d} \mathcal{N}(0, \sigma_\varphi^2)$$

To prove (iii), it is sufficient to show that

$$\begin{aligned}\hat{\sigma}_\varphi^2 &:= \frac{1}{M\hat{p}(s_{\text{obs}})} \sum_{i=1}^M (\varphi(\theta_i) - \widehat{\mathbb{E}_\pi(\varphi)})^2 \hat{w}^2(\theta_i, u_i) \\ &\xrightarrow{a.s.} \frac{\mathbb{E}_{\tilde{g}_{\text{IS}}}(X^2)}{p(s_{\text{obs}})^2} = \sigma_\varphi^2.\end{aligned}$$

□

References

- Andrieu, C. and Roberts, G. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37:697–725.
- Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.
- Jacob, P. and Thiery, A. H. (2015). On non-negative unbiased estimators. *Annals of Statistics*, 43(2):769–784.
- Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.
- McLeish, D. (2012). A general method for debiasing a Monte Carlo estimator. *Monte Carlo Methods and Applications*, 17:301–315.
- Moller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient Markov Chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.
- Rhee, C. H. and Glynn, P. W. (2013). Unbiased estimation with square root convergence for SDE model. Technical report, Stanford University.
- Sisson, S. A. and Fan, Y. (2011). Likelihood-free Markov chain Monte Carlo. In Brooks, S. P., Gelman, A., Jones, G., and Meng, X.-L., editors, *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC Press.
- Tavare, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring Coalescence Times From DNA Sequence Data. *Genetics*, 145(2):505–518.
- Tran, M.-N., Scharth, M., Pitt, M. K., and Kohn, R. (2013). Importance sampling squared for Bayesian inference in latent variable models. <http://arxiv.org/abs/1309.3339>.